U.S. Census Bureau

Monthly Retail Trade and Food Services Technical Documentation

SAMPLE DESIGN AND ESTIMATION PROCEDURES

The U.S. Census Bureau introduced new samples with the 2005 Annual Retail Trade Survey (ARTS) and the September 2006 Monthly Retail Trade Survey (MRTS). The new samples are designed to produce estimates based on the 2002 North American Industry Classification System (NAICS). This section describes the design, selection, and estimation procedures for the new samples. For descriptions of the prior samples see the *Annual Revision of Monthly Retail and Food Services (formerly called the Annual Benchmark Report for Retail Trade*, or prior benchmark reports.

Sampling Frame

The sampling frame used for the MRTS and the ARTS (ARTS) has two types of sampling units represented—Employer Identification Numbers (EINs) and large, multiple-establishment firms. Both sampling units represent clusters of one or more establishments owned or controlled by the same firm. The information used to create these sampling units was extracted from data collected as part of the 2002 Economic Census and from establishment records contained on the Census Bureau's Business Register as updated to December 2004. The next few paragraphs give details about the Business Register; the distinction between firms, EINs, and establishments; and the construction of the sampling units. Though important, they are not essential to understanding the basic sample design and readers may continue to the **Stratification, Sampling Rates, and Allocation** section.

The Business Register is a multirelational database that contains a record for each known establishment that is located in the United States or one of its territories and has employees. An *establishment* is a single physical location where business transactions take place and for which payroll and employment records are kept. Groups of one or more establishments under common ownership or control are *firms*. A *single-unit* firm owns or operates only one establishment. A *multiunit* firm owns or operates two or more establishments. The treatment of establishments on the Business Register differs according to whether the establishment is part of a *single-unit* or *multiunit* firm. In particular, the structure

of an establishment's primary identifier on the Business Register differs depending on whether it is owned by a *single-unit* firm or by a *multiunit* firm.

A single-unit firm's primary identifier is its EIN. The Internal Revenue Service (IRS) issues the EIN and the firm uses it as an identifier to report social security payments for its employees under the Federal Insurance Contributions Act (FICA). The same act requires all employer firms to use EINs. Each employer firm is associated with at least one EIN and only one firm can use a given EIN. Because a single-unit firm has only one establishment, there is a one-to-one relationship between the firm and the EIN. Thus the firm, the EIN, and the establishment all reference the same physical location and all three terms can be used interchangeably and unambiguously when referring to a single-unit firm.

For multiunit firms, however, a different structure connects the firm with its establishments via the EIN. Essentially, a multiunit firm is associated with a cluster of one or more EINs and EINs are associated with one or more establishments. A multiunit firm consists of at least two establishments. Each firm is associated with at least one EIN and only one firm can use a given EIN. However, one multiunit firm may have several EINs. Similarly, there is a one-to-many relationship between EINs and establishments. Each EIN can be associated with many establishments, but each establishment is associated with only one EIN. Because of the possibility of one-to-many relationships, we must distinguish between the firm, its EINs, and its establishments. The multiunit firm that owns or controls a particular establishment is identified on the Business Register by way of the establishment's primary identifier.

The primary identifier of an establishment owned by a multiunit firm consists of a unique combination of an alpha number and a plant number. The alpha number identifies the multiunit firm, and the plant number identifies a particular establishment within that firm. All establishments owned or controlled by the same multiunit firm have the same alpha number. Different multiunit firms have different alpha numbers, and different establishments within the same multiunit firm have different plant numbers. The Census Bureau assigns both the alpha number to the multiunit firm and plant numbers to the corresponding establishments based on the results of the quinquennial economic census and the annual Company Organization Survey.

To create the sampling frame, we extract the records for all employer establishments located in the United States and classified in the Retail Trade and Accommodation and Food Services sectors as defined by the 2002 NAICS. For these establishments, we extract sales, payroll, employment, name and address information, as well as primary identifiers and, for establishments owned by multiunit firms, associated EINs.

To create the sampling units for multiunit firms, we aggregate the economic data of the establishments owned by these firms to an EIN level by tabulating the

establishment data for all inscope establishments associated with the same EIN. Similarly, we aggregate the data to a multiunit firm level by tabulating the establishment data for all inscope establishments associated with the same alpha number. No aggregation is necessary to put single-unit establishment information on an EIN basis or a firm basis. Thus, the sampling units created for single-unit firms simultaneously represent establishment, EIN, and firm information. In summary, the sampling frame is a complex amalgam of establishments, EINs, and firms.

Stratification, Sampling Rates, and Allocation

The primary stratification of the sampling frame is by industry group based on the detail required for publication. We further stratify the sampling units within industry group by a measure of size (substratify) related to their annual revenue. Sampling units expected to have a large effect on the precision of the estimates are selected ''with certainty.'' This means they are sure to be selected and will represent only themselves (i.e., have a selection probability of 1 and a sampling weight of 1). Within each industry stratum, we determine a substratum boundary (or cutoff) that divides the certainty units from the noncertainty units. We base these cutoffs on a statistical analysis of data from the 2002 Economic Census. Accordingly, these values are on a 2002 sales basis. We also use this analysis to determine the number of size substrata for each industry stratum and to set preliminary sampling rates needed to achieve specified sampling variability constraints on revenue estimates for different industry groups. The size substrata and sampling rates are later updated through analysis of the sampling frame.

Sample Selection

The first step in the sample selection identified firms selected with certainty. If a firm's annual sales or end-of-year inventories were greater than the corresponding certainty cutoff, that firm was selected into the sample with certainty. The MRTS and ARTS samples use the same certainty firms.

All firms not selected with certainty were subjected to sampling on an EIN basis. If a firm had more than one EIN, we treated each of its EINs as a separate sampling unit. To be eligible for the initial sampling, an EIN had to have nonzero payroll in 2003. The EINs were stratified according to their major industry and their estimated revenue (on a 2002 basis). Within each noncertainty stratum, a simple random sample of EINs was selected without replacement. The selected noncertainty EINs are divided into two equal groups. One group is canvassed for both the monthly and the annual survey, the other group is canvassed for only the annual survey. Therefore, the noncertainty sample for the annual survey is twice the size of the noncertainty sample for the monthly sample.

Quarterly Birth Sampling

Periodically, we update the samples to represent new EINs appearing on the Business Register. These new EINs, called *births*, are EINs recently assigned by the IRS on the latest available IRS mailing list for FICA taxpayers and assigned an industry classification (if possible) by the Social Security Administration (SSA).

EIN births are sampled on a quarterly basis using a two-phase selection procedure. To be eligible for selection, a birth must either have no industry classification or be classified in an industry within the scope of the ARTS, the Annual Wholesale Trade Survey (AWTS) , or the Service Annual Survey (SAS), and it must meet certain criteria regarding its number of paid employees or quarterly payroll. In the first phase, births are stratified by broad industry groups and a measure of size based on quarterly payroll or expected number of paid employees. The birth is assigned to the payroll or employment stratum with the larger sampling fraction. This procedure is conservative because it results in the birth being assigned the smaller of two possible first phase sampling weights. A relatively large sample is selected using equal probability systematic sampling. The selected births are canvassed to obtain a more reliable measure of size, consisting of sales in 2 recent months, company affiliation information, and a new or more detailed industry classification code. Births that have not returned their questionnaire after 30 days are contacted by telephone.

Using this more reliable information, the selected births from the first phase are subjected to probability proportional-to-size sampling with overall probabilities equivalent to those used in drawing the initial MRTS and ARTS samples from the December 2004 Business Register. Because of the time it takes for a new employer firm to acquire an EIN from the IRS, and because of the time needed to accomplish the two-phase birth-selection procedure, births are added to the samples approximately 9 months after they begin operation. The births selected for the MRTS sample are a subset of the births selected for the ARTS sample.

If a firm was selected with certainty and had more than one establishment at the time of sampling, any new establishments that the firm acquires, even if under new or different EINs, are included in the sample with certainty. However, if a single-unit firm was selected with certainty, only future establishments associated with that firm's originally-selected EIN are included in the sample with certainty; any new EINs that might later be associated with that firm are subjected to sampling through the quarterly birth-selection procedure.

Single-unit EINs selected into the sample with certainty are not dropped from canvass and tabulation if they are no longer on the IRS mailing list. Rather, the firm that used the EIN is contacted, and if a successor EIN is found, it is added to the survey. For both inactive and reactivated EINs, data are tabulated for only the portion of the reference year that these EINs reported payroll to the IRS.

Sample Maintenance Procedures for the Annual Survey

Births that are selected in the quarterly birth-selection procedure in November of the annual survey reference year are included in the initial mailing of the annual survey questionnaires in January of the following year. To better represent all EIN births in the reference year, and specifically to account for the lag between the time a business starts operation and the time it takes to acquire an EIN and identify and select the EIN into one of our surveys, we add births to the annual survey sample that are selected in February, May, and August of the year following the annual survey reference year. We mail annual survey forms to these births in June and August to supplement the initial annual survey mailings.

Sample Maintenance Procedures for the Monthly Survey

Because births are not represented in the monthly survey until they go through the two-phase selection procedure, an interim procedure is used to account for births during the period between the onset of activity and the time of birth selection. This interim procedure consists of imputing data for all EINs currently in the monthly survey that go out of business but are still on the IRS mailing list.

Births are added to the monthly survey in February, May, August, and November of each year. At the same time, deaths are removed from the survey. To minimize the effect of births and deaths on the month-to-month change estimates, we phase-in these changes by incrementally increasing the sampling weights of the births and decreasing the sampling weights of the deaths in a similar fashion. In the first month, we tabulate the births at one-third their sampling weight and tabulate the deaths at two-thirds their sampling weight. In the second month, we tabulate the births at two-thirds their sampling weight and tabulate the deaths at one-third their sampling weight. In the third month, we tabulate the births at their full sampling weight and the deaths are dropped (sampling weight equal zero).

Procedures for Producing Monthly Estimates

Estimates of monthly sales and end-of-month inventories are derived from data collected in the MRTS. Each month, firms in the MRTS sample are asked to report their sales and inventory data for the month just ending. Monthly totals are computed as the sum of weighted data (reported and imputed) for all selected sampling units that meet the sample canvass and tabulation criteria given below. The weight for a given sampling unit is the reciprocal of its probability of selection into the sample. The monthly totals are then benchmarked to the latest available annual survey totals. See the Revisions to Previously Published Estimates section for a description of the benchmarking procedures.

To be eligible for the sample canvass and tabulation, an EIN selected in the noncertainty sampling operations must meet both of the following requirements:

• It must be on the latest available IRS mailing list for FICA taxpayers from the previous quarter.

• It must have been selected from the Business Register in either the initial sampling or during the quarterly birth-selection procedure.

Monthly total estimates for broad industry groups (e.g., two-, three-, and four-digit NAICS levels) are computed by summing the benchmarked monthly totals for the appropriate detailed industries comprising the broader industry group.

Variances are estimated using the method of random groups.

Procedures for Producing Annual Estimates

Estimates of annual sales and end-of-year inventories are derived from data collected in the ARTS. Firms in the ARTS sample are asked to report their sales and inventory data for the year just ending. Two years of data are requested in the year in which a new sample is introduced. Annual totals are computed as the sum of weighted data (reported and imputed) for all selected sampling units that meet the sample canvass and tabulation criteria given above. The weight for a given sampling unit is the reciprocal of its probability of selection into the ARTS sample. The annual estimates are adjusted using results of the 2002 Economic Census. See <http://www.census.gov/svsd/www/summary.html> for a description of the adjustment procedures.

Annual total estimates for broad industry groups (e.g., two-, three-, and four-digit NAICS levels) are computed by summing the census-adjusted annual totals for the appropriate detailed industries comprising the broader industry group.

Variances are estimated using the method of random groups.

Reliability of Estimates

The published estimates may differ from the actual, but unknown, population values. For a particular estimate, statisticians define this difference as the total error of the estimate. When describing the accuracy of survey results, it is convenient to discuss total error as the sum of sampling error and nonsampling error. Sampling error is the error arising from the use of a sample, rather than a census, to estimate population values. Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate. The sampling error of an estimate can usually be estimated from the sample, whereas the nonsampling error of an estimate is difficult to measure and can rarely be estimated. Consequently, the actual error in an estimate exceeds the error that can be estimated. Further descriptions of sampling error and nonsampling error are provided in the following sections. Data users should take into account the estimates of sampling error and the potential effects of nonsampling error when using the published estimates.

Sampling Error

Because the estimates are based on a sample, exact agreement with results that would be obtained from a complete enumeration of firms on the sampling frame using the same enumeration procedures is not expected. However, because each firm on the sampling frame has a known probability of being selected into the sample, it is possible to estimate the sampling variability of the survey estimates.

The particular sample used in this survey is one of a large number of samples of the same size that could have been selected using the same design. If all possible samples had been surveyed under the same conditions, an estimate of a population parameter of interest could have been obtained from each sample. For the parameter of interest, estimates derived from the different samples would, in general, differ from each other. Common measures of the variability among these estimates are the sampling variance, the standard error, and the coefficient of variation (CV). The sampling variance is defined as the squared difference, averaged over all possible samples of the same size and design, between the estimator and its average value. The standard error is the square root of the sampling variance. The CV expresses the standard error as a percentage of the estimate to which it refers. For example, an estimate of 200 units that has an estimated standard error of 10 units has an estimated CV of 5 percent. The sampling variance, standard error, and CV of an estimate can be estimated from the selected sample because the sample was selected using probability sampling. Note that measures of sampling variability, such as the standard error and CV, are estimated from the sample and are also subject to sampling variability. (Technically, we should refer to the *estimated* standard error or the *estimated* CV of an estimator. However, for the sake of brevity we have omitted this detail.) It is important to note that the standard error and CV only measure sampling variability. They do not measure any systematic biases in the estimates.

The estimate from a particular sample and its associated standard error can be used to construct a confidence interval. A *confidence interval* is a range about a given estimator that has a specified probability of containing the average of the estimates for the parameter derived from all possible samples of the same size and design. Associated with each interval is a percentage of confidence, which is interpreted as follows. If, for each possible sample, an estimate of a population parameter and its approximate standard error were obtained, then:

1. For approximately 90 percent of the possible samples, the interval from 1.65 standard errors below to 1.65 standard errors above the estimate would include the average of the estimates derived from all possible samples of the same size and design.

2. For approximately 95 percent of the possible samples, the interval from 1.96 standard errors below to 1.96 standard errors above the estimate would include

the average of the estimates derived from all possible samples of the same size and design.

To illustrate the computation of a confidence interval for an estimate of total revenue, assume that an estimate of total revenue is $10,750 million and the CV for this estimate is 1.8 percent, or 0.018. First, obtain the standard error of the estimate by multiplying the total revenue estimate by its CV. For this example, multiply $10,750 million by 0.018. This yields a standard error of $193.5 million. The upper and lower bounds of the 90-percent confidence interval are computed as $10,750 million ± 1.65 x $193.5 million. Consequently, the 90-percent confidence interval is $10,431 million to $11,069 million. If corresponding confidence intervals were constructed for all possible samples of the same size and design, approximately 9 out of 10 (90 percent) of these intervals would contain the average of the estimates derived from all possible samples.

The Census Bureau recommends that individuals using published estimates incorporate this information into their analyses, as sampling error could affect the conclusions drawn from these estimates.

Nonsampling Errors

Nonsampling error encompasses all other factors, other than sampling error, that contribute to the total error of a sample survey estimate and may also occur in censuses. It is often helpful to think of nonsampling error as arising from deficiencies or mistakes in the survey process. Nonsampling errors are difficult to measure and can be attributed to many sources: the inclusion of erroneous units in the survey (overcoverage), the exclusion of eligible units from the survey (undercoverage), nonresponse, misreporting, mistakes in recording and coding responses, misinterpretation of questions, and other errors of collection, response, coverage, or processing. Although nonsampling error is not measured directly, the Census Bureau employs quality control procedures throughout the process to minimize this type of error.

A potential source of bias in the estimates is nonresponse. Nonresponse is defined as the inability to obtain all the intended measurements or responses about all selected units. Two types of nonresponse are often distinguished. *Unit nonresponse* is used to describe the inability to obtain any of the substantive data for a sampled unit. In most cases of unit nonresponse, the questionnaire was never returned to the Census Bureau after several attempts to elicit a response. *Item nonresponse* occurs either when a question is unanswered or the response to the question fails computer or analyst edits.

For both unit and item nonresponse, a missing value is replaced by a predicted value obtained from an appropriate model for nonresponse. This procedure is called *imputation* and uses survey data and administrative data as input. In any given month, imputed data amount to about **22** percent of the total monthly retail

and food services sales estimate and about **28** percent of the total retail end-of-month inventory estimate. For the annual survey, imputed data amount to about **6** percent of the total retail sales and food services estimate and about **6** percent of the total retail end-of-year inventory estimate.

Source: Retail Indicators Branch, U.S. Census Bureau
Last Revised: April 30, 2008